

Adaptive kernel estimation of the baseline function in the Cox model, with high-dimensional covariates

Agathe Guilloux

Laboratoire de Statistique Théorique et Appliquée,

Université Pierre et Marie Curie - Paris 6

e-mail : `agathe.guilloux@upmc.fr`

Sarah Lemler

Laboratoire de Mathématiques et Modélisation d'Évry, UMR CNRS 8071- USC INRA,

Université d'Évry Val d'Essonne, France

e-mail : `sarah.lemler@genopole.cnrs.fr`

Marie-Luce Taupin

Laboratoire de Mathématiques et Modélisation d'Évry, UMR CNRS 8071- USC INRA,

Université d'Évry Val d'Essonne, France

e-mail : `marie-luce.taupin@genopole.cnrs.fr`

Unité MaIAGE, INRA Jouy-En-Josas, France

Abstract

The aim of this article is to propose a novel kernel estimator of the baseline function in a general high-dimensional Cox model, for which we derive non-asymptotic rates of convergence. To construct our estimator, we first estimate the regression parameter in the Cox model via a Lasso procedure. We then plug this estimator into the classical kernel estimator of the baseline function, obtained by smoothing the so-called Breslow estimator of the cumulative baseline function. We propose and study an adaptive procedure for selecting the bandwidth, in the spirit of Goldenshluger and Lepski (2011). We state non-asymptotic oracle inequalities for the final estimator, which reveal the reduction of the rates of convergence when the dimension of the covariates grows.

Keywords: Cox's proportional hazards model; Conditional hazard rate function; Semi-parametric model; High-dimensional covariates; Counting processes; Kernel estimation; Goldenshluger and Lepski method, Non-asymptotic oracle inequalities; Survival analysis

1 Introduction

The Cox model, introduced by [Cox \(1972\)](#), is a regression model often considered in survival analysis to relate the distribution of a time T to the values of covariates. The hazard function of T is then defined by

$$\lambda_0(t, \mathbf{Z}) = \alpha_0(t) \exp(\boldsymbol{\beta}_0^T \mathbf{Z}), \quad (1)$$

where $\mathbf{Z} = (Z_1, \dots, Z_p)^T$ is a p -dimensional vector of covariates, $\boldsymbol{\beta}_0 = (\beta_{0_1}, \dots, \beta_{0_p})^T$ the vector of regression coefficients and α_0 the baseline hazard function.

The regression parameter $\boldsymbol{\beta}_0$ and the baseline function α_0 are the two unknown parameters in this model. Yet, more attention has been paid to the estimation of the regression parameter than to the estimation of the baseline function.

There are good reasons for this. First, the Cox partial log-likelihood, introduced by [Cox \(1972\)](#), allows to estimate $\boldsymbol{\beta}_0$ without the knowledge of α_0 . Secondly, the regression parameter is directly related to the covariates. Therefore, in order to select the relevant covariates that explain the best the survival time, we need to estimate the regression parameter. A lot of papers deal with the problem of the estimation of $\boldsymbol{\beta}_0$, the number of covariates p being large or not compared with the size of the panel n . When p is smaller than n , the usual estimator of $\boldsymbol{\beta}_0$ is obtained by maximizing the Cox partial log-likelihood (see [Andersen et al. \(1993\)](#) as a reference book). When the number of covariates grows, the Lasso procedure is often considered. This procedure consists in the minimization of the opposite of the ℓ_1 -penalized Cox partial log-likelihood. Asymptotic results are stated in [Bradic et al. \(2012\)](#), [Kong and Nan \(2012\)](#), [Bradic and Song \(2012\)](#). Finally, the non-asymptotic rate of convergence of the Lasso is now known to be of order $\sqrt{\log p/n}$, see [Huang et al. \(2013\)](#).

The estimation of the baseline function α_0 has been less studied. The known estimator of the baseline function is a kernel estimator, introduced by [Ramlau-Hansen \(1983a;b\)](#). We present here its form in the special case of right-censoring. Let us consider, for the moment, that we observe for $i = 1, \dots, n$, $(X_i, \delta_i, \mathbf{Z}_i)$, where $X_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$, T_i is the time of interest and C_i the censoring time. The usual kernel estimator is then obtained from an estimator of the cumulative baseline function A_0 defined by $A_0(t) = \int_0^t \alpha_0(s) ds$. This estimator is called the Breslow estimator and is defined, for $t > 0$, by

$$\hat{A}_0(t, \hat{\boldsymbol{\beta}}) = \sum_{i=1}^n \frac{\delta_i}{S_n(X_i, \hat{\boldsymbol{\beta}})} \quad \text{with } S_n(t, \hat{\boldsymbol{\beta}}) = \sum_{i: T_i \geq t} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{Z}_i), \quad (2)$$

see [Ramlau-Hansen \(1983b\)](#) and [Andersen et al. \(1993\)](#) for details. From $\hat{A}_0(\cdot, \hat{\boldsymbol{\beta}})$, the kernel function estimator for α_0 is derived by smoothing the increments of the Breslow estimator. It is defined by

$$\hat{\alpha}_{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{h} \int_0^\tau K\left(\frac{t-u}{h}\right) d\hat{A}_0(u, \hat{\boldsymbol{\beta}}), \quad \tau \geq 0 \quad (3)$$

with $K : \mathbb{R} \mapsto \mathbb{R}$ a kernel with integral 1, and h a positive parameter called the bandwidth. This estimator has been introduced and studied by [Ramlau-Hansen \(1983a;b\)](#) within the framework of the multiplicative intensity model for counting processes, thereby extending its use to censored survival data. Consistency and asymptotic normality are proven in [Ramlau-Hansen \(1983b\)](#) with fixed bandwidth.

The choice of the bandwidth in kernel estimation is crucial, in particular when one is interested in establishing non-asymptotic adaptive inequalities. State-of-the-art methods are based on cross-validation. [Ramlau-Hansen \(1981\)](#) has suggested the cross-validation method to select the bandwidth but without any theoretical guarantees. For randomly censored survival data, [Marron and Padgett \(1987\)](#) have shown that the cross-validation method gives the optimal bandwidth for estimating the density: the ratio between the integrated squared error for the cross-validation bandwidth and the infimum of the integrated squared error for any bandwidth almost surely converges to one. [Grégoire \(1993\)](#) has considered the cross-validated method suggested by [Ramlau-Hansen \(1981\)](#) for the adaptive estimation of the intensity of a counting process and has proved some consistency and asymptotic normality results for the cross-validated kernel estimator. However, all the results for the adaptive kernel estimator with a cross-validated bandwidth are asymptotic.

No non-asymptotic oracle inequalities have to date been stated for the kernel estimator of the baseline function. In addition, to our knowledge, the construction of $\hat{\alpha}_h^{\hat{\beta}}$ has not yet been considered for high-dimensional covariates. The objective of the present paper is then twofold: whatever the dimension, we aim at proposing an estimator $\hat{\alpha}^{\hat{\beta}}$ of the baseline function, for which we can establish a non-asymptotic oracle inequality to measure its performances. The loss of prediction of $|\hat{\alpha}^{\hat{\beta}} - \alpha_0|$ when p increases will be quantified.

To fulfill these two purposes, the idea is to estimate first the regression parameter β_0 via a Lasso procedure applied to the Cox partial log-likelihood, then to plug this estimator in the usual kernel estimator (3) of the baseline hazard function and finally to select a data-driven bandwidth, following a procedure adapted from [Goldenshluger and Lepski \(2011\)](#). In the latter, the problem of bandwidth selection in kernel density estimation is addressed and an adaptive estimator is derived, which satisfies non-asymptotic minimax bounds. This method was then considered by [Doumic et al. \(2012\)](#) to estimate the division rate of a size-structured population in a non-parametric setting, by [Bouaziz et al. \(2013\)](#) to estimate the intensity function of a recurrent event process and by [Chagny \(2014\)](#) for the estimation of a real function via a warped kernel strategy. In the present paper, we consider it to obtain an adaptive kernel estimator of the baseline function with a data-driven bandwidth. We establish the first adaptive and non-asymptotic oracle inequality, which warrants the theoretical performances of this kernel estimator. The oracle inequality depends on the non-asymptotic control of $|\hat{\beta} - \beta_0|_1$ deduced from an estimation inequality stated by [Huang et al. \(2013\)](#) and extended to the case of unbounded counting processes (see [Guilloux et al. \(2015\)](#) for details).

The paper is organized as follows. In Section 3, we describe the two-step procedure to estimate the baseline function: first, we describe the estimation of β_0 as a preliminary step and give the bound for $|\hat{\beta} - \beta_0|_1$ and then we focus on the kernel estimation of α_0 and describe the adaptive estimation procedure of Goldenshluger and Lepski to select a

data-driven bandwidth. In Section 4, we establish a non-asymptotic oracle inequality for the adaptive kernel estimator. The fundamental proofs are gathered in Section 6. Lastly, a supplementary material provides some technical results needed in the proofs.

2 Notations and preliminaries

2.1 Framework with counting processes

Consider the general setting of counting processes, which embeds the classical case of right censoring. We follow here the now classical setting of Andersen et al. (1993) or Fleming and Harrington (2011). For n independent individuals, we observe for $i = 1, \dots, n$ a counting process N_i , a random process Y_i with values in $[0, 1]$ and a vector of covariates $\mathbf{Z}_i = (Z_{i,1}, \dots, Z_{i,p})^T \in \mathbb{R}^p$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t \geq 0}$ be the filtration defined by

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), 0 \leq s \leq t, \mathbf{Z}_i, i = 1, \dots, n\}.$$

From the Doob-Meyer decomposition, we know that each N_i admits a compensator denoted by Λ_i , such that $M_i = N_i - \Lambda_i$ is a $(\mathcal{F}_t)_{t \geq 0}$ local square-integrable martingale (see Andersen et al. (1993) for details). We assume in the following that N_i satisfies an Aalen multiplicative intensity model.

Assumption 2.1. *For each $i = 1, \dots, n$ and all $t \geq 0$,*

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \mathbf{Z}_i) Y_i(s) ds, \quad (4)$$

where $\lambda_0(t, \mathbf{z}) = \alpha_0(t) e^{\beta^T \mathbf{z}}$, for $\mathbf{z} \in \mathbb{R}^p$.

This general setting, introduced by Aalen (1980), embeds several particular examples as censored data, marked Poisson processes and Markov processes (see Andersen et al. (1993) for further details). This framework generalizes the case considered in Ramlau-Hansen (1983b) to unbounded counting processes and hence widens the scope of applications: we can consider the jumps of the counting to happen at times of relapse from a disease in biomedical research, times of monetization in marketing, times of blogging in social network study, etc.

2.2 Notations

For a real number $q \geq 1$ and a function $f : \mathbb{R} \mapsto \mathbb{R}$ such that $|f|^q$ is integrable and bounded, we consider

$$\|f\|_{L^q(\mathbb{R})} = \left(\int_{\mathbb{R}} |f(x)|^q dx \right)^{1/q} \text{ and } \|f\|_{\infty} = \sup_{x \in \mathbb{R}} |f(x)|.$$

The integrals and the supremum are restricted to the support of f and for τ a positive real number, we set $\|f\|_{\infty, \tau} = \sup_{x \in [0, \tau]} |f(x)|$ and we simply denote by $\|\cdot\|_2$ the \mathbb{L}^2 -norm restricted to the interval $[0, \tau]$, so that

$$\|f\|_2^2 = \int_0^\tau f^2(x) dx.$$

For h a positive real number, we define $f_h(\cdot) = f(\cdot/h)/h$. For square-integrable functions f and g from \mathbb{R} to \mathbb{R} , we denote the convolution product of f and g by $f * g$. For a vector $\mathbf{b} \in \mathbb{R}^p$ and a real $q \geq 1$, we denote $|\mathbf{b}|_q = (\sum_{j=1}^p |b_j|^q)^{1/q}$.

For quantities $\gamma(n)$ and $\eta(n)$, the notation $\gamma(n) \lesssim \eta(n)$ means that there exists a positive constant c such that $\gamma(n) \leq c\eta(n)$.

Finally, let $\mathbf{Z} \in \mathbb{R}^p$ denote the generic vector of covariates with the same distribution as the vectors of covariates \mathbf{Z}_i of each individual i and by Z_j its j -th component, namely the j -th covariates of the vector \mathbf{Z} .

3 Estimation procedure

In this section, we describe the two-step procedure to estimate the baseline function. We begin by recalling the usual estimation of the regression parameter β_0 in high-dimension. We then focus our study on the second step, which consists in the adaptive kernel estimation of the baseline function α_0 .

3.1 Preliminary estimation of β_0

The regression parameter β_0 is estimated via a Lasso procedure applied to the so-called Cox partial log-likelihood introduced by [Cox \(1972\)](#) and defined, for all $\beta \in \mathbb{R}^p$, by

$$l_n^*(\beta) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \log \frac{e^{\beta^T \mathbf{Z}_i}}{S_n(t, \beta)} dN_i(t), \quad \text{where } S_n(\beta, t) = \frac{1}{n} \sum_{i=1}^n e^{\beta^T \mathbf{Z}_i} Y_i(t). \quad (5)$$

The estimator $\hat{\beta}$ of β_0 is then defined by

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{B}(0, R)} \{-l_n^*(\beta) + \text{pen}(\beta)\}, \quad \text{with } \text{pen}(\beta) = \Gamma_n |\beta|_1, \quad (6)$$

where Γ_n is a positive regularization parameter to be suitably chosen and $\mathcal{B}(0, R)$ is the ball defined by

$$\mathcal{B}(0, R) = \{b \in \mathbb{R}^p : |b|_1 \leq R\}, \quad \text{with } R > 0.$$

The ball constraint has already been considered by [van de Geer \(2008\)](#) or [Kong and Nan \(2012\)](#). Roughly speaking, it means that we have restrict our attention to a, possibly very large, ball around β_0 , for which the following (very mild) assumption is needed. It is required to control the kernel estimator of the baseline function β_0 .

Assumption 3.1. *We assume that $|\beta_0|_1 < +\infty$.*

Concerning the covariates, we introduce the following assumption.

Assumption 3.2. *There exists a positive constant B such that for all $j \in \{1, \dots, p\}$,*

$$|Z_j| \leq B.$$

Assumption 3.2 is a classical assumption in the Cox model to obtain oracle inequalities (see Huang et al. (2013) and Bradic and Song (2012)) and seems reasonable since in practice

We now give a general version of the estimation inequality of Theorem 3.1 of Huang et al. (2013). We refer to Guillaux et al. (2015) for a proof of Proposition 3.3 in the general case.

Proposition 3.3. *Let $k > 0$, $c > 0$ and $s := \text{Card}\{j \in \{1, \dots, p\} : \beta_{0j} \neq 0\}$ be the sparsity index of β_0 . Assume that $\|\alpha_0\|_{\infty, \tau} < \infty$. Then, under Assumptions 3.1 and 3.2, with probability larger than $1 - cn^{-k}$, we have*

$$|\hat{\beta} - \beta_0|_1 \leq C(s) \sqrt{\frac{\log(pn^k)}{n}} \quad (7)$$

where $C(s) > 0$ is a constant depending on the sparsity index s .

In the rest of the paper, the conditions of Proposition 3.3 will be fulfilled, so that $\hat{\beta}$ satisfies Inequality (7). The assumption $\|\alpha_0\|_{\infty, \tau} < \infty$ is to be found in Assumptions 3.4.

3.2 Estimation of α_0

In this subsection, we define the kernel estimator of the baseline hazard function α_0 on which our procedure relies. We state some functional and kernel assumptions, and we describe the Goldenshluger and Lepski procedure to select a data-driven bandwidth.

3.2.1 Kernel estimator

We first recall the definition of the kernel estimator introduced by Ramlau-Hansen (1983b) by using kernel functions to smooth the increments of the non-parametric Breslow estimator (2) of the cumulative intensity.

Let define $K : \mathbb{R} \rightarrow \mathbb{R}$ a kernel, namely K is a function such that $\int_{\mathbb{R}} K(x) dx = 1$. The usual kernel function estimator of α_0 is then defined by

$$\hat{\alpha}_h^{\hat{\beta}}(t) = \frac{1}{nh} \sum_{i=1}^n \int_0^{\tau} K\left(\frac{t-u}{h}\right) \frac{\mathbb{1}_{\{\bar{Y}(u) > 0\}}}{S_n(u, \hat{\beta})} dN_i(u), \quad (8)$$

with

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \text{and} \quad S_n(u, \beta) = \frac{1}{n} \sum_{i=1}^n e^{\beta^T Z_i} Y_i(u), \quad \text{for all } \beta \in \mathbb{R}^p.$$

The parameter $h > 0$ is called the bandwidth. In kernel function estimation, the bandwidth has to be chosen by the user. Grégoire (1993) has defined a cross-validation procedure for selecting the bandwidth for the smooth estimate of intensity in the Aalen counting process. To our knowledge, all theoretical results for the kernel function estimator (8) with a bandwidth selected by cross-validation are asymptotic. The cross-validation ensures no theoretical adaptive guarantees when the size of the panel n is fixed and not so large as it is the case for medical surveys where only a few patients can be observed. This explains our interest in providing a data-driven method to select automatically the bandwidth and obtain a kernel function estimator, for which we can warrant some non-asymptotic properties.

In what follows, we denote the estimator under study by $\hat{\alpha}_h^{\hat{\beta}}$ in which the Lasso estimator (6) has been plugged.

3.2.2 Functional and kernel assumptions

Classical conditions are required on the intensity function and the kernel K .

Assumption 3.4.

- (i) For all $i \in \{1, \dots, n\}$, the random process Y_i takes its values in $\{0, 1\}$.
- (ii) For $S(t, \beta_0) = \mathbb{E}[e^{\beta_0^T Z_i} Y_i(t)]$, there exists a positive constant c_S such that,

$$S(t, \beta_0) \geq c_S, \quad \forall t \in [0, \tau].$$

- (iii) $\|\alpha_0\|_{\infty, \tau} := \sup_{t \in [0, \tau]} \alpha_0(t) < \infty$.

Assumption 3.4.(i) is satisfied for all the examples quoted in the introduction. In fact, this assumption is needed to ensure that the random process Y_i has a lower bound when it is nonzero. We could also have considered a modified estimator of $S_n(u, \beta)$, defined by (5), as it is usually done in the censoring case without covariates. Assumption 3.4.(ii) is common in the context of estimation with censored observations (see Andersen et al. (1993)). Assumption 3.4.(iii) is required to obtain Lemma 6.1 and Theorem 4.1 below. Nevertheless, the value $\|\alpha_0\|_{\infty, \tau}$ is not needed to compute the estimator (see Section 5).

The following assumptions are fulfilled by many standard kernel functions and are standard in kernel function estimation.

Assumption 3.5.

- (i) $\|K\|_{\infty} = \sup_{u \in \mathbb{R}} |K(u)| < \infty$ and $\|K\|_2^2 = \int_{\mathbb{R}} K^2(u) du < \infty$.
- (ii) $nh \geq 1$ and $0 < h < 1$.
- (iii) The kernel K is of order 1, i.e. for $j \in \{0, 1, 2\}$ the function $x \mapsto x^j K(x)$ is integrable and

$$\int_{\mathbb{R}} x K(x) dx = 0 \quad \text{and} \quad \int_{\mathbb{R}} x^2 K(x) dx < \infty.$$

Assumptions 3.5.(i) and 3.5.(ii) are rather standard in kernel density estimation (see Goldenshluger and Lepski (2011)) and has also been considered in the kernel intensity estimation by Bouaziz et al. (2013). Assumption 3.5.(iii) is only required to ensure that $K_h * \alpha_0(t) \xrightarrow{h \rightarrow 0} \alpha_0(t)$ for all $t \in [0, \tau]$.

Remark 3.6. *In this paper, we do not assume that the kernel K has a compact support, by opposition to Bouaziz et al. (2013). The Breslow estimator (8) and the pseudo-estimator (19) are then well defined for all $t \in [0, \tau]$.*

3.2.3 Collection of estimators

Let \mathcal{H}_n be a grid of bandwidths $h > 0$, satisfying the following assumptions:

Assumption 3.7.

(i) $\text{Card}(\mathcal{H}_n) \leq n$.

(ii) For some $a \geq 0$, $\sum_{h \in \mathcal{H}_n} \frac{1}{nh} \lesssim \log^a(n)$.

(iii) For all $b > 0$, $\sum_{h \in \mathcal{H}_n} \exp(-b/h) < +\infty$.

Assumptions 3.7.(i)-(iii) mean that the bandwidth collection should not be too large. Let us give an example of grid \mathcal{H}_n that satisfies the three previous assumptions.

Example 3.8 (Example of \mathcal{H}_n). *The following grid is considered in the simulations in Section 5*

$$\mathcal{H}_n = \left\{ h_j = \frac{1}{2^j}, j = 1, \dots, \lfloor \log(n)/\log(2) \rfloor \right\},$$

where $\varepsilon \in [0, \tau/2]$ is a small constant chosen arbitrarily as close as possible to 0. For this grid, all the assumptions required on the bandwidths are verified. Indeed, $\text{Card}(\mathcal{H}_n) \leq \log(n)/\log(2) \leq n$ and $\forall k = 1, \dots, \lfloor \log(n)/\log(2) \rfloor$, we have $h_j \in [n^{-1}, 1]$. Moreover, Assumption 3.7.(ii) holds true since

$$\sum_{j: h_j \in \mathcal{H}_n} \frac{1}{nh_j} = \frac{1}{n} \sum_{j=1}^{\lfloor \log(n)/\log(2) \rfloor} 2^j = O(1).$$

Lastly,

$$\sum_{j: h_j \in \mathcal{H}_n} \exp(-b/h_j) = \sum_{j=1}^{\lfloor \log(n)/\log(2) \rfloor} e^{-b2^j} = O(1)$$

and Assumption 3.7.(iii) is verified.

On the grid \mathcal{H}_n , we obtain a set of kernel estimators $\mathcal{F}(\mathcal{H}_n) = \{\hat{\alpha}_h^{\hat{\beta}}, h \in \mathcal{H}_n\}$ of the baseline function α_0 from the definition (8).

3.2.4 Adaptive selection of the bandwidth

We wish to automatically select a relevant bandwidth $h \in \mathcal{H}_n$, in such a way to then be able to select a kernel estimator among the set $\mathcal{F}(\mathcal{H}_n)$. As usual, we must choose a bandwidth h which realizes the best compromise between the squared-bias and the variance terms. The choice should be data-driven. For this, we use a quite recent method introduced by [Goldenshluger and Lepski \(2011\)](#) for the problem of density estimation. The "Goldenshluger and Lepski method" has only been considered in two different settings: [Bouaziz et al. \(2013\)](#) has applied this method to provide an adaptive kernel function estimator of the intensity function of a recurrent event process and [Chagny \(2014\)](#) has used it to estimate a real valued function from a sample of random couples (see [Chagny \(2014\)](#)). Lately, [Chagny \(2013\)](#) has also proposed a "mixed strategy", which consists in applying the "Goldenshluger and Lepski method" to select the relevant model in model selection methods for real valued function in regression models. We consider this method to obtain an adaptive kernel function estimator of the baseline function, for which we establish a non-asymptotic oracle inequality.

Let us begin to describe the method. We can explain the idea of the method of [Goldenshluger and Lepski \(2011\)](#) from an heuristic proposed by [Chagny \(2013\)](#). We want to define $\hat{\alpha}_{h\hat{\beta}}^{\hat{\beta}}$ so that the risk is as close as possible as

$$\min_{h \in \mathcal{H}_n} \{ \|\alpha_0 - K_h * \alpha_0\|_2^2 + V(h) \},$$

with

$$V(h) = \kappa \frac{\|\alpha_0\|_{\infty, \tau}^\tau}{c_S^2} \left(\|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]_\tau + \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] \right) \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh} = O\left(\frac{1}{nh}\right),$$

for a constant $\kappa > 0$. In order to get closer from the bias term $\|\alpha_0 - K_h * \alpha_0\|_2^2$, we replace α_0 with an estimator $\hat{\alpha}_{h'}^{\hat{\beta}}$ (with a fixed bandwidth h'), so that we obtain $\|\hat{\alpha}_{h'}^{\hat{\beta}} - K_h * \hat{\alpha}_{h'}^{\hat{\beta}}\|_2^2$. However, unlike the bias term, this quantity is random and thus contains some variability. We need to correct this variability by deducting the part of the variance $V(h')$. Lastly, since there are no reason to choose one bandwidth $h' \in \mathcal{H}_n$ rather than an other one, we consider the entire collection and take the maximum over this collection.

Formally, we define for $h \in \mathcal{H}_n$

$$A^{\hat{\beta}}(h) = \sup_{h' \in \mathcal{H}_n} \left\{ \|\hat{\alpha}_{h, h'}^{\hat{\beta}} - \hat{\alpha}_{h'}^{\hat{\beta}}\|_2^2 - V(h') \right\}_+ \quad (9)$$

where

$$\hat{\alpha}_{h, h'}^{\hat{\beta}}(t) = K_{h'} * \hat{\alpha}_h^{\hat{\beta}}(t), \quad (10)$$

for any $t \geq 0$ and h, h' two positive real numbers, and

$$V(h) = \kappa \frac{\|\alpha_0\|_{\infty, \tau}^\tau}{c_S^2} \left(\|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]_\tau + \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] \right) \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh}, \quad (11)$$

for some numerical constant $\kappa > 0$. A data-driven equivalent of this variance term is given in Section 5. The choice of κ is also discussed.

From these definitions, we deduce the following choice of the bandwidth:

$$\hat{h}^{\hat{\beta}} = \arg \min_{h \in \mathcal{H}_n} \{A^{\hat{\beta}}(h) + V(h)\}. \quad (12)$$

Our adaptive kernel estimator is then $\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}}$.

4 Non-asymptotic bounds for the kernel estimator

Now, let us state the main theorems of the chapter, which provide the first non-asymptotic oracle inequality for the adaptive kernel baseline estimator in high-dimension.

Theorem 4.1. *Under Assumptions 3.1, 3.2, 3.4.(i)-(iii) and 3.5.(i)-(iii), if \mathcal{H}_n is a finite discrete set of bandwidths such that 3.7.(i)-(iii) are satisfied, then there exists a constant κ such that $\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}}$ defined by (11), (9) and (12) satisfies for n large enough and $k \geq 12$:*

$$\mathbb{E}[\|\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \alpha_0\|_2^2] \leq C \inf_{h \in \mathcal{H}_n} \left\{ \|\alpha_h - \alpha_0\|_2^2 + V(h) \right\} + C'(s) \frac{\log^a(n) \log(pn^k)}{n}, \quad (13)$$

$$(14)$$

with

$$V(h) = \kappa \frac{\|\alpha_0\|_{\infty, \tau} \tau}{c_S^2} \left(\|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]_{\tau} + \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] \right) \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh},$$

where C is a numerical constant, $C'(s)$ a constant depending on τ , κ_b , B , $|\beta_0|_1$, R , $\|\alpha_0\|_{\infty, \tau}$, c_S , $\|K\|_{\mathbb{L}^1(\mathbb{R})}$, $\|K\|_{\mathbb{L}^2(\mathbb{R})}$ and on the sparsity index s of β_0 .

This inequality ensures that the adaptive kernel estimator $\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}}$ automatically makes the squared-bias/variance compromise. The selected bandwidth $\hat{h}^{\hat{\beta}}$ is performing as well as the unknown oracle:

$$h_{oracle}^{\hat{\beta}} := \arg \min_{h \in \mathcal{H}_n} \mathbb{E}[\|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0\|_2^2],$$

up to the multiplicative constant C and up to a remaining term of order $\log^a(n) \log(pn^k)/n$, which is negligible. In Inequality (16), the infimum term is classic in such oracle inequalities for kernel estimators: the bias term $\|\alpha_h - \alpha_0\|_2^2$ decreases when h decreases and the variance term $V(h)$ increases when h decreases. The remaining terms are of order

$$\frac{\log^a(n) \log(pn^k)}{n} = \frac{k \log^{a+1}(n)}{n} + \frac{\log^a(n) \log(p)}{n}.$$

Chagny (2014), in the context of an additive regression model, has established an oracle inequality for the kernel estimator of the real-value regression function with a remaining term of order $1/n$. In the context of the estimation of the intensity of a recurrent event

process observed under a standard censoring scheme but without covariates, [Bouaziz et al. \(2013\)](#) have a logarithm term which appears in their oracle inequality with a remaining term of order $\log^{1+a}(n)/n$ instead of the expected $1/n$. This logarithm term comes from the control in $\log(n)/n$ between the distribution function G and its modified Kaplan-Meier estimator \hat{G} , which appears in the kernel intensity estimator. The exponent a in the remaining term arises from Assumption 3.7.(ii), which is needed for the control of the difference between the kernel intensity estimator involving \hat{G} and a pseudo-estimator that does not depend of \hat{G} . As well as [Bouaziz et al. \(2013\)](#), our kernel estimator depends on an other estimator, so that we need Assumption 3.7.(ii) in order to control the difference between the kernel estimator (8) and the pseudo-estimator (19). If our kernel estimator had not involved another estimator, we would have considered condition $\sum 1/h \leq k_0 n_0^a$, as in [Chagny \(2014\)](#), instead of Assumption 3.7.(ii). The term in $\log(p)/n$ in the remaining term comes from the control of $|\hat{\beta} - \beta_0|_1$ given by Proposition 3.3. This term is typical of the estimation of the regression parameter β_0 when the number of covariates is large. There is no hope of capturing up to usual rates in this high-dimensional setting, but the loss in the variance term is only of order $\log p/n$.

When we assume that the counting processes N_i are bounded for $i = 1, \dots, n$, the variance term $V(h)$ is simpler and has the same form as the variance term in [Bouaziz et al. \(2013\)](#). In this particular case, Theorem 4.1 takes the following form.

Theorem 4.2. *Under the same assumptions as in Theorem 4.1 and assuming also that there exists $c_\tau > 0$, such that $N_i(t) \leq c_\tau$ almost surely for every $t \in [0, \tau]$ and $i \in \{1, \dots, n\}$, there exists a constant κ such that $\hat{\alpha}_{h\hat{\beta}}^{\hat{\beta}}$ defined by (9), (12) and*

$$V_b(h) = \kappa \frac{c_\tau \tau \|\alpha_0\|_{\infty, \tau}}{c_S} \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh}, \quad (15)$$

satisfies for n large enough:

$$\mathbb{E}[\|\hat{\alpha}_{h\hat{\beta}}^{\hat{\beta}} - \alpha_0\|_2^2] \leq \tilde{C} \inf_{h \in \mathcal{H}_n} \left\{ \|\alpha_h - \alpha_0\|_2^2 + V_b(h) \right\} + \tilde{C}'(s) \frac{\log^a(n) \log(np)}{n} \quad (16)$$

$$(17)$$

where \tilde{C} is a numerical constant, $\tilde{C}'(s)$ a constant depending on τ , c_τ , B , $|\beta_0|_1$, R , $\|\alpha_0\|_{\infty, \tau}$, c_S , $\|K\|_{\mathbb{L}^1(\mathbb{R})}$, $\|K\|_{\mathbb{L}^2(\mathbb{R})}$ and on the sparsity index s of β_0 .

The proof of Theorem 4.2 is close to the one of Theorem 4.1 and we refer to [Lemler \(2014\)](#) for the details.

5 Applications

5.1 Simulation study

The aim of this subsection is to illustrate the behavior of the kernel estimator $\hat{\alpha}_{h\hat{\beta}}^{\hat{\beta}}$ of the baseline function in the case of right censoring and to compare it with the usual kernel

estimator with a bandwidth selected by cross-validation introduced by [Ramblau-Hansen \(1983b\)](#).

5.1.1 Simulated datas: censored data.

We consider a cohort of size n and p covariates simulated according to the Cox model (1) with right censoring and with regression parameter β_0 chosen as a vector of dimension p , defined by

$$\beta_0 = (0.1, 0.3, 0.5, 0, \dots, 0)^T \in \mathbb{R}^p,$$

for various $p \geq 3$. Several choices of n and p have been considered, with sample size n taking values $n = 200$ and $n = 500$ and p varying between $p = \sqrt{n}$, being 15 and 22 respectively and $p = n$, referred to as the high-dimension case. For each n and p , the design matrix $\mathbf{Z} = (Z_{i,j})_{1 \leq i \leq n, 1 \leq j \leq p}$ is simulated independently from a uniform distribution on $[-1, 1]$ and survival times $T_i, i = 1, \dots, n$ are simulated according to Weibull distributions $\mathcal{W}(1.5, 1)$ and $\mathcal{W}(0.5, 2)$. Hence, the associated baseline function has the form $\alpha_0(t) = a\lambda^a t^{a-1}$, where a and λ stand for parameters in $\mathcal{W}(a, \lambda)$. The censoring times C_i , for $i = 1, \dots, n$, are simulated independently from the survival times via an exponential distribution $\mathcal{E}(1/\gamma \mathbb{E}[T_1])$, where γ is adjusted to the chosen rate of censorship: $\gamma = 4.5$ for 20% of censorship and $\gamma = 1.2$ for 50% of censorship.

The time τ of the end of the study is taken as the quantile at 90% of $(T_i \wedge C_i)_{i=1, \dots, n}$. For $i = 1, \dots, n$, we compute the observed times $X_i = \min(T_i, \tilde{C}_i)$, where $\tilde{C}_i = C_i \wedge \tau$ and the censoring indicators $\delta_i = \mathbb{1}_{T_i \leq C_i}$. The definition of \tilde{C}_i ensures that there exist some $i \in \{1, \dots, n\}$ for which $X_i \geq \tau$, so that all estimators are defined on the interval $[0, \tau]$ and it prevents from a certain edge effect. Each sample $(\mathbf{Z}_i, T_i, C_i, X_i, \delta_i, i = 1, \dots, n)$ is repeatedly simulated $N_e = 100$ times.

The compared estimators of the baseline hazard function are both constructed with the Epanechnikov kernel, defined by

$$K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}_{\{|u| \leq 1\}}.$$

In both cases we plugg the Lasso regression parameter estimator $\hat{\beta}$ defined by (6) and implemented from the R-package *glmnet*.

We compare two procedures for the data-driven choice of h : the Goldenshluger and Lepski method with the selected bandwidth denoted by $\hat{h}_{GL}^{\hat{\beta}}$ and the cross-validation with the selected bandwidth denoted by $\hat{h}_{CV}^{\hat{\beta}}$.

5.1.2 The Goldenshluger and Lepski method

The adaptive bandwidth selection method, we consider here, is based on the grid of bandwidths \mathcal{H}_n defined in Example 3.8 by

$$\mathcal{H}_n = \{1/2^k, k = 0, \dots, \lfloor \log(n)/\log(2) \rfloor\}.$$

In our procedure (9), the variance term $V(h)$ involves unknown quantities, so we consider a data-driven equivalent of it and use that the right-censoring context implies that the counting processes are bounded. Hence we implement the following procedure:

$$\hat{h}_{GL}^{\hat{\beta}} = \arg \min_{h \in \mathbb{H}_n} \{A^{\hat{\beta}}(h) + \hat{V}_b^{\hat{\beta}}(h)\},$$

where, for any $t \geq 0$ and h, h' two positive real numbers,

$$A^{\hat{\beta}}(h) = \sup_{h' \in \mathcal{H}_n} \left\{ \|\hat{\alpha}_{h,h'}^{\hat{\beta}} - \hat{\alpha}_{h'}^{\hat{\beta}}\|_2^2 - \hat{V}_b^{\hat{\beta}}(h') \right\}_+, \quad \hat{\alpha}_{h,h'}^{\hat{\beta}}(t) = K_{h'} * \hat{\alpha}_h^{\hat{\beta}}(t),$$

and

$$\hat{V}_b^{\hat{\beta}}(h) = \kappa' \frac{\|\hat{\alpha}_{\max(h)}^{\hat{\beta}}\|_{\infty, \tau} \|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh}.$$

In the variance term $\hat{V}_b^{\hat{\beta}}(h)$, we have replaced the true unknown function α_0 by an estimator $\hat{\alpha}_{\max(h)}^{\hat{\beta}}$ computed for the largest bandwidth h in the grid \mathcal{H}_n (see Bouaziz et al. (2013)). The numerical constant κ' is a universal constant that we tuned from the comparison of the MISEs for several candidate values in the range $10^{-4} - 1000$, and for the two different distributions $\mathcal{W}(1.5, 1)$ and $\mathcal{W}(0.5, 2)$. We take $\kappa' = 1$.

5.1.3 Cross-validation method

The bandwidth $\hat{h}_{CV}^{\hat{\beta}}$ selected by cross-validation is defined by:

$$\hat{h}_{CV}^{\hat{\beta}} = \arg \min_h \left\{ \mathbb{E} \int_0^\tau (\hat{\alpha}_h^{\hat{\beta}}(t))^2 dt - 2 \sum_{i \neq j} \frac{1}{h} K\left(\frac{X_i - X_j}{h}\right) \frac{\delta_i}{\bar{Y}(X_i)} \frac{\delta_j}{\bar{Y}(X_j)} \right\},$$

where $\bar{Y} = \sum_{i=1}^n \mathbb{1}_{\{X_i \geq t\}}$.

5.1.4 Performances

The performances of these two estimators are evaluated via different Integrated Squared Errors (ISE). For some function $\alpha \in \mathbb{L}^2([0, \tau])$ the standard ISE and the total ISE are respectively defined by

$$\begin{aligned} \text{ISEstand}(\alpha) &= \int_0^\tau (\alpha(t) - \alpha_0(t))^2 dt, \\ \text{ISEtotal}(\alpha, \beta) &= \frac{1}{n} \sum_{i=1}^n \int_0^\tau (\alpha(t) e^{\beta^T \mathbf{Z}_i} - \alpha_0(t) e^{\beta_0^T \mathbf{Z}_i})^2 dt. \end{aligned}$$

The associated Mean Integrated Squared Errors are defined by $\text{MISEg}(\alpha) = \mathbb{E}[\text{ISEg}(\alpha)]$, for g=stand or total, where the expectation is taken on (T_i, C_i, \mathbf{Z}_i) (for sake of simplicity, we write $\text{MISEg}(\alpha)$ even if the MISE depends on β). We obtain an estimation of the different MISE by taking the empirical mean for the $N_e = 100$ replications.

Dimensions \ MISEs		20%				50%			
		MISEstand		MISEtot		MISEstand		MISEtot	
$n = 200$	$p = 15$	0.014	0.017	0.080	0.082	0.023	0.029	0.104	0.120
	$p = 500$	0.013	0.016	0.117	0.117	0.022	0.026	0.152	0.154
$n = 500$	$p = 22$	0.009	0.007	0.038	0.035	0.011	0.012	0.055	0.056
	$p = 1000$	0.008	0.008	0.068	0.064	0.011	0.013	0.094	0.096

Table 1: MISEs of the kernel estimators with a bandwidth selected by the Goldenshluger and Lepski method (first column for each MISE) and with a bandwidth selected by cross-validation of the baseline function with a Lasso estimator of the regression parameter, given two rates of censoring: 20% and 50% of censoring.

5.1.5 Results

Table 1 gives the two empirical MISEs of the kernel estimators with a bandwidth selected either by cross-validation or by the Goldenshluger and Lepski method for a Lasso estimator of the regression parameter and survival times that are distributed from $\mathcal{W}(1.5, 1)$, in different censoring situation. We consider the results for two rates of censoring: a usual rate of 20% of censoring and large rate of 50% of censoring.

As expected, with both procedures, the MISEs are degraded when the censoring rate increases. When we compare the standard and total MISEs, the results are rather good for the standard MISE. This is consistent, since the total MISE measures the performances of the complete intensity estimators $\hat{\lambda}(t, \mathbf{Z}) = \hat{\alpha}^{\hat{\beta}}(t)e^{\hat{\beta}^T \mathbf{Z}}$, including the error coming from $\hat{\beta}$, whereas the standard MISEs measures the performances of the estimators of the baseline function. Therefor

One can see that MISEs resulting from the two procedure are very similar, with very rather good results with our procedure.

In Table 2, we give the standard MISE of the kernel estimators with a bandwidth selected either by cross-validation or by the Goldenshluger and Lepski method for different distributions of the survival times. We observe that the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method performs better than the one with a bandwidth selected by cross-validation for the two Weibull distributions.

5.2 Application to a real dataset on breast cancer

In this section, we apply the proposed method to study the relapse free survival (RFS) from breast cancer adjusted on high-dimensional covariates in two groups of patients. We consider a Cox model (1) to link the RFS to the covariates. We aim at answering the two questions of the introduction concerning the biomarkers that influence the RFS and the prediction of the RFS for each individual. The dataset is available on the website www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532.

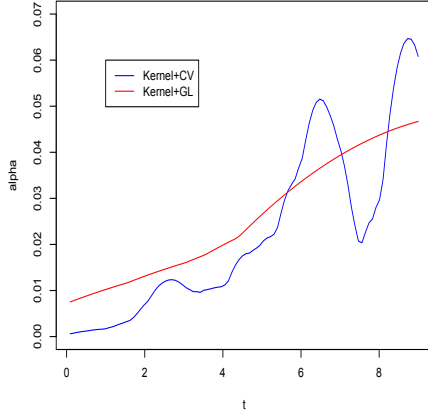
Distributions		$\mathcal{W}(1.5, 1)$		$\mathcal{W}(0.5, 2)$	
Dimensions					
$n = 200$	$p = 15$	0.056	0.088	1.02	1.561
	$p = 200$	0.06	0.085	0.923	1.556
$n = 500$	$p = 22$	0.025	0.037	1.006	1.521
	$p = 500$	0.027	0.033	1.098	1.515

Table 2: MISEs for the kernel estimators with a bandwidth selected by the Goldenshluger and Lepski method (first column for each distribution) and with a bandwidth selected by cross-validation (second column for each distribution), with a Lasso estimator of the regression parameter for two different Weibull distributions of the survival times.

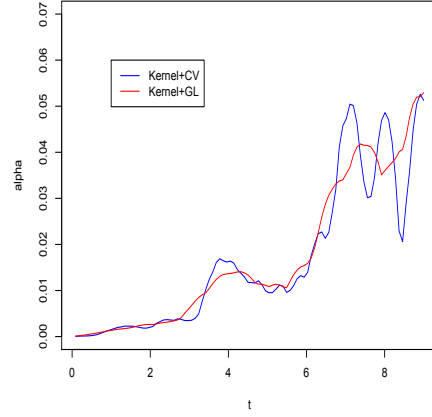
The dataset consists of 414 patients in the cohort GSE6532 collected by [Loi et al. \(2007\)](#) for the purpose of characterizing Estrogen Receptor (ER)-positive subtypes with gene expression profiles. Estrogen receptors are a group of proteins found inside cells, which is activated by the hormone estrogen. There are different forms of estrogen receptors, referred to as subtypes of estrogen receptors. When they are over expressed, they are referred to as ER-positive. The dataset has been studied from a survival analysis point of view in [Tian et al. \(2012\)](#). Following them, we apply the two procedures to the same survival time of interest (the RFS). Excluding patients with incomplete informations, as it is done by [Loi et al. \(2007\)](#), there are 142 patients receiving Tamoxifen and 104 untreated patients. It should be underlined that we should do better to handle the missing data, but in this study we also exclude the patients with missing data. In addition to clinical informations such as the age or the size of the tumor, we have 44 928 gene expression measurements for each of the 246 patients. Two different survival times are available in this study: the time of relapse free survival and the time of distant metastasis free survival. We are interested in this study in the time of relapse free survival, which subjects to right censoring due to incomplete follow-up. There are 60% of censorship in the group of the untreated patients and 66% in the group of patients receiving Tamoxifen. Our goal is to compare the baseline functions in the two groups of patients: the patients receiving Tamoxifen and the untreated patients.

We start by a preliminary variable selection among the 44928 levels of gene expression. This corresponds to a screening step (see [Fan et al. \(2010\)](#)). This preliminary variable selection is based on the score statistics of each Cox model considered for each variable separately. We only keep the variables which score statistics are superior to a threshold. The difference from the procedure proposed by [Fan et al. \(2010\)](#) is that we fix the number of covariates we want to keep and then we tune a threshold to select this number of covariates. We define the threshold as the 95th percentile of a Chi-squared distribution with 1 degree of freedom, so that 996 probesets have been selected and with the clinical covariates, we have $p = 1000$.

Figure 1 shows the graphs of the kernel estimators of the baseline function with a bandwidth selected by cross-validation and by the Goldenshluger and Lepski method, in



(a) Untreated patients ($p=1000$).



(b) Tamoxifen patients ($p=1000$).

Figure 1: Kernel estimator with a bandwidth selected by cross-validation (in blue) and kernel estimator with a bandwidth selected with the Goldenshluger and Lepski method (in red). The righthand plot is associated to the group of untreated patients and the lefthand plot correspond to the group of Tamoxifen patients for $p = 1000$.

the two groups of patients for $p = 1000$.

On Figure 1, we observe that the estimator obtained by cross-validation fails to give an interpretable estimate of α_0 for the untreated patients. For the estimator obtained from the Goldenshluger and Lepski method, we observe that the risk of relapse to breast cancer has slowed down with the treatment, because the estimated baseline function is close to 0 until $t = 2.5$ for the patients treated with tamoxifen whereas it already increases at time $t = 1.5$ for the untreated patients. This leads us to believe that the treatment has a positive influence on the survival time.

6 Proofs

This section is organized as follows. First, we establish a lemma that allows to control the estimation error of the kernel estimator for a fixed bandwidth h , then we prove Theorem 4.1 from two fundamental lemmas that are also proved in this section. We add a supplementary material for all the other used technical lemmas, that are not essential for a first reading.

6.1 Intermediate lemma: bound for the kernel estimator of α_0 with a fixed bandwidth

We first establish a non-asymptotic global bound on Mean Integrated Squared Error (MISE) for the estimators $\hat{\alpha}_h^{\hat{\beta}}$, with h fixed.

Lemma 6.1. *Under Assumptions 3.1, 3.2, 3.4.(ii)-(iii) and 3.5.(i)-(iii), for a fixed $h > 0$, n large enough and $k \geq 12$*

$$\mathbb{E}[|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0|_2^2] \leq 2\|\alpha_h - \alpha_0\|_2^2 + \frac{C_1}{nh} + C_2(s) \frac{\log(pn^k)}{n} \quad (18)$$

where C_1 is a constant depending on τ , $\|\alpha_0\|_{\infty, \tau}$, c_S , $\mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}]$, $\mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]$, τ and $\|K\|_{\mathbb{L}^2(\mathbb{R})}$ and $C_2(s)$ is a constant depending on B , $|\beta_0|_1$, R , $\|\alpha_0\|_{\infty, \tau}$, c_S , τ , $\|K\|_{\mathbb{L}^2(\mathbb{R})}$ and on the sparsity index s of β_0 .

To prove this lemma and link the kernel estimator to the true baseline function α_0 , the trick is to introduce a pseudo-estimator, which does not depend on $\hat{\beta}$. Consider for $h > 0$ the pseudo-estimator

$$\bar{\alpha}_h(t) = \frac{1}{nh} \sum_{i=1}^n \int_0^\tau K\left(\frac{t-u}{h}\right) \frac{1}{S(u, \beta_0)} dN_i(u), \quad (19)$$

which corresponds to the kernel estimator of α_0 when $S(u, \beta_0) = \mathbb{E}[e^{\beta_0^T \mathbf{Z}_i} Y_i(u)]$ is known. To justify the choice of the pseudo-estimator, let us calculate its expectation:

$$\begin{aligned} \mathbb{E}[\bar{\alpha}_h(t)] &= \frac{1}{nh} \sum_{i=1}^n \int_0^\tau K\left(\frac{t-u}{h}\right) \frac{1}{S(u, \beta_0)} \alpha_0(u) \mathbb{E}[e^{\beta_0^T \mathbf{Z}_i} Y_i(u)] du \\ &= \frac{1}{h} \int_0^\tau K\left(\frac{t-u}{h}\right) \alpha_0(u) du \\ &= K_h * \alpha_0(t), \end{aligned}$$

which is a unit approximation of α_0 , so that $K_h * \alpha_0 \xrightarrow{h \rightarrow 0} \alpha_0$ under mild conditions (see Bochner Lemma and Assumption 3.5.(iii)).

In the following, we define for all $t \in [0, \tau]$

$$\alpha_h(t) := \mathbb{E}[\bar{\alpha}_h(t)] = K_h * \alpha_0(t). \quad (20)$$

The proof is based on the following decomposition for $h > 0$

$$\mathbb{E}[|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0|_2^2] \leq 2\mathbb{E}[|\hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h|_2^2] + 2\mathbb{E}[|\bar{\alpha}_h - \alpha_0|_2^2]. \quad (21)$$

Since the pseudo-estimator (19) does not depend on the estimator $\hat{\beta}$, the error $\mathbb{E}[|\bar{\alpha}_h - \alpha_0|_2^2]$ is easier to bound than directly the error $\mathbb{E}[|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0|_2^2]$. The study of the error of $\hat{\alpha}_h^{\hat{\beta}} - \alpha_0$ is then divided into two parts: the error of $\bar{\alpha}_h - \alpha_0$ and the one of $\hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h$.

The following lemma provides the classical bias/variance inequality for the pseudo-estimator (19).

Lemma 6.2. Under Assumptions 3.4.(ii)-(iii), 3.5.(i)-(iii), for $h > 0$ fixed

$$\mathbb{E}[\|\bar{\alpha}_h - \alpha_0\|_2^2] \leq \|\alpha_h - \alpha_0\|_2^2 + \frac{2\|\alpha_0\|_{\infty,\tau}\tau}{c_S^2} \left(\mathbb{E}[e^{\beta_0^T \mathbf{z}_1}] + \|\alpha_0\|_{\infty,\tau} \mathbb{E}[e^{2\beta_0^T \mathbf{z}_1}] \tau \right) \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh}. \quad (22)$$

The next lemma controls the quadratic error between $\hat{\alpha}_h^{\hat{\beta}}$ and $\bar{\alpha}_h$. The term to be controlled in this difference is in fact the difference between the regression parameter β_0 and its Lasso estimator $\hat{\beta}$. The ℓ_1 -norm of this difference is bounded from Proposition 3.3 by a term of order $\log(np)/n$. This explains the obtained bound in the following lemma.

Lemma 6.3. Under Assumptions 3.4.(ii)-(iii), 3.5.(i)-(iii), 3.1 and 3.2, for a fixed $h > 0$,

$$\mathbb{E}[\|\hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h\|_2^2] \leq c(s) \frac{\log(n^k p)}{n},$$

where $c(s)$ is a constant depending on B , $|\beta_0|_1$, R , $\|\alpha_0\|_{\infty,\tau}$, c_S , τ , $\|K\|_{\mathbb{L}^2(\mathbb{R})}$ and s the sparsity index of β_0 .

From Equation (21), gathering Lemmas 6.2 and 6.3 provide directly Lemma 6.1.

Lemmas 6.2 and 6.3 are proved in the supplementary material.

6.2 Proof of the oracle inequality in Theorem 4.1

For all $h \in \mathcal{H}_n$, $A^{\hat{\beta}}(h)$ is defined by (9) and we can apply this definition for $h = \hat{h}^{\hat{\beta}}$. We deduce from this, using Definition (12) of $\hat{h}^{\hat{\beta}}$, that for all $h \in \mathcal{H}_n$

$$\begin{aligned} \|\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \alpha_0\|_2^2 &\leq 3\|\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \hat{\alpha}_{h,\hat{h}^{\hat{\beta}}}^{\hat{\beta}}\|_2^2 + 3\|\hat{\alpha}_{h,\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \hat{\alpha}_h^{\hat{\beta}}\|_2^2 + 3\|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0\|_2^2 \\ &\leq 3(A^{\hat{\beta}}(h) + V(\hat{h}^{\hat{\beta}})) + 3(A^{\hat{\beta}}(\hat{h}^{\hat{\beta}}) + V(h)) + 3\|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0\|_2^2 \\ &\leq 6(A^{\hat{\beta}}(h) + V(h)) + 3\|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0\|_2^2. \end{aligned}$$

We obtain for $h \in \mathcal{H}_n$

$$\mathbb{E}[\|\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \alpha_0\|_2^2] \leq 6\mathbb{E}[A^{\hat{\beta}}(h)] + 6V(h) + 3\mathbb{E}[\|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0\|_2^2]. \quad (23)$$

Lemma 6.1 gives a bound of $\mathbb{E}[\|\hat{\alpha}_h^{\hat{\beta}} - \alpha_0\|_2^2]$, which reveals the bias term, the variance term of order $1/nh$ and a remaining term of order $\log(np)/n$, and $V(h)$ is of the expected order $1/nh$. $\mathbb{E}[A^{\hat{\beta}}(h)]$ is bounded in the following proposition.

Proposition 6.4. Let $h \in \mathcal{H}_n$ be fixed. Under the assumptions of Theorem 4.1, there exist constants C_1 , $C_2(s)$, $C_3(s)$ such that,

$$\mathbb{E}[A^{\hat{\beta}}(h)] \leq C_1\|\alpha_h - \alpha_0\|_2^2 + C_2(s) \frac{\log^a(n) \log(n^k p)}{n} + C_3(s) \frac{\log(n^k p)}{n}, \quad (24)$$

where the constant C_1 only depends on $\|K\|_1$.

Applying Inequalities (18) and (24) in Equation (23) implies Inequality (16) by taking the infimum over $h \in \mathcal{H}_n$. This ends the proof of Theorem 4.1. \square

6.3 Proof of Proposition 6.4

We introduce several additional notations $\bar{\alpha}_{h,h'} = K_{h'} * \bar{\alpha}_h$, $\alpha_h(t) = \mathbb{E}[\bar{\alpha}_h(t)]$, $\alpha_{h,h'}(t) = \mathbb{E}[\bar{\alpha}_{h,h'}(t)]$, and write

$$\begin{aligned} A^{\hat{\beta}}(h) &= \sup_{h' \in \mathcal{H}_n} \left\{ \|\hat{\alpha}_{h'}^{\hat{\beta}} - \hat{\alpha}_{h,h'}^{\hat{\beta}}\|_2^2 - V(h') \right\}_+ \\ &\leq 5 \sup_{h' \in \mathcal{H}_n} \left\{ \|\bar{\alpha}_{h'} - \alpha_{h'}\|_2^2 - V(h')/10 \right\}_+ + 5 \sup_{h' \in \mathcal{H}_n} \left\{ \|\bar{\alpha}_{h,h'} - \alpha_{h,h'}\|_2^2 - V(h')/10 \right\}_+ \\ &\quad + 5 \sup_{h' \in \mathcal{H}_n} \|\hat{\alpha}_{h'}^{\hat{\beta}} - \bar{\alpha}_{h'}\|_2^2 + 5 \sup_{h' \in \mathcal{H}_n} \|\hat{\alpha}_{h,h'}^{\hat{\beta}} - \bar{\alpha}_{h,h'}\|_2^2 + 5 \sup_{h' \in \mathcal{H}_n} \|\alpha_{h'} - \alpha_{h,h'}\|_2^2 \\ &:= 5(T_1 + T_2 + T_3 + T_4 + T_5) \end{aligned}$$

- Study of $\mathbb{E}[T_1]$: Recall that for all $h \in \mathcal{H}_n$

$$\|\bar{\alpha}_h - \alpha_h\|_2^2 = \sup_{f \in \mathbb{L}^2([0, \tau]), \|f\|_2=1} \langle \bar{\alpha}_h - \alpha_h, f \rangle_2^2. \quad (25)$$

We introduce the centered empirical process $\nu_{n,h}(f) = \langle \bar{\alpha}_h - \alpha_h, f \rangle_2$, which is equal to

$$\frac{1}{n} \sum_{i=1}^n \int_0^\tau f(t) \left(\int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} (dN_i(u) - \alpha_0(u)S(u, \beta_0)du) \right) dt.$$

As $f \mapsto \nu_{n,h}(f)$ is continuous, the supremum in (25) can be taken over a countable dense subset of $\{f \in \mathbb{L}^2([0, \tau]), \|f\|_2 = 1\}$, which we denote by \mathcal{B}_τ . Therefore, we can write

$$\begin{aligned} \mathbb{E}[T_1] &\leq \mathbb{E} \left[\left\{ \sup_{h' \in \mathcal{H}_n} \|\bar{\alpha}_{h'} - \alpha_{h'}\|_2^2 - V(h')/10 \right\}_+ \right] \\ &\leq \sum_{h' \in \mathcal{H}_n} \mathbb{E} \left[\left\{ \|\bar{\alpha}_{h'} - \alpha_{h'}\|_2^2 - V(h')/10 \right\}_+ \right] \\ &\leq \sum_{h' \in \mathcal{H}_n} \mathbb{E} \left[\left\{ \sup_{f \in \mathcal{B}_\tau} \nu_{n,h'}^2(f) - V(h')/10 \right\}_+ \right]. \end{aligned} \quad (26)$$

Let introduce a key lemma, which allows to bound (26).

Lemma 6.5. *Let us introduced the centered process $\nu_{n,h}(f) = \langle \bar{\alpha}_h - \alpha_h, f \rangle_2$, for any $h \in \mathcal{H}_n$ and $f \in \mathbb{L}^2([0, \tau])$ and $\mathcal{B}_\tau = \{f \in \mathbb{L}^2([0, \tau]), \|f\|_2 = 1\}$. Under the assumptions of Theorem 4.1, with $V(h')$ defined by (11) for any $h' \in \mathcal{H}_n$, there exists two constants c_6 and c_7 depending on the bound κ_b of the B r kholder Inequality, τ , $\|\alpha_0\|_{\infty, \tau}$, the bound c_S of $S(t, \beta_0)$, $\mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}]$, $\mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]$, $\|K\|_{\mathbb{L}^1(\mathbb{R})}$ and $\|K\|_{\mathbb{L}^2(\mathbb{R})}$, such that*

$$\sum_{h \in \mathcal{H}_n} \mathbb{E} \left[\left\{ \sup_{f \in \mathcal{B}_\tau(h)} \nu_{n,h}^2(f) - V(h)/10 \right\}_+ \right] \leq \frac{c_6}{n} + c_7 \frac{\log^a(n)}{n}.$$

So, from Lemma 6.5, there exists two constants $c_6 > 0$ and $c_7 > 0$ such that

$$\mathbb{E}[T_1] \leq \frac{c_6}{n} + c_7 \frac{\log^a(n)}{n}, \quad (27)$$

where c_6 and c_7 depend on τ , $\|\alpha_0\|_{\infty, \tau}$, c_S , $\mathbb{E}[e^{\beta_0^T Z_1}]$, $\mathbb{E}[e^{2\beta_0^T Z_1}]$, $\|K\|_{\mathbb{L}^1(\mathbb{R})}$ and $\|K\|_{\mathbb{L}^2(\mathbb{R})}$.

• Study of $\mathbb{E}[T_2]$: We study T_2 similarly as T_1 since

$$\mathbb{E}[T_2] \leq \sum_{h' \in \mathcal{H}_n} \mathbb{E} \left[\left\{ \|\bar{\alpha}_{h,h'} - \alpha_{h,h'}\|_{2,h'}^2 - V(h')/10 \right\}_+ \right].$$

From Lemma 6.5 (see the remark at the end of the proof of Lemma 6.5), there exists two constants $c_8 > 0$ and $c_9 > 0$ such that

$$\mathbb{E}[T_2] \leq \frac{c_8}{n} + c_9 \frac{\log^a(n)}{n}, \quad (28)$$

where c_8 and c_9 depend on τ , $\|\alpha_0\|_{\infty, \tau}$, c_S , $\mathbb{E}[e^{\beta_0^T Z_1}]$, $\mathbb{E}[e^{2\beta_0^T Z_1}]$, $\|K\|_{\mathbb{L}^1(\mathbb{R})}$ and $\|K\|_{\mathbb{L}^2(\mathbb{R})}$.

• Study of $\mathbb{E}[T_3]$: First, write for all $h \in \mathcal{H}_n$, that

$$\hat{\alpha}_h^{\hat{\beta}}(t) - \bar{\alpha}_h(t) = \frac{1}{nh} \sum_{i=1}^n \int_0^\tau K\left(\frac{t-u}{h}\right) \frac{S(u, \beta_0) \mathbb{1}_{\{\bar{Y}(u) > 0\}} - S_n(u, \hat{\beta})}{S_n(u, \hat{\beta}) S(u, \beta_0)} dN_i(u)$$

For all $u \in [0, \tau]$, we have $S(u, \beta_0) \mathbb{1}_{\{\bar{Y}(u) > 0\}} - S_n(u, \hat{\beta}) = (S(u, \beta_0) - S_n(u, \hat{\beta})) \mathbb{1}_{\{\bar{Y}(u) > 0\}}$. Indeed, for all $u \in [0, \tau]$, if $\mathbb{1}_{\{\bar{Y}(u) > 0\}} = 0$, then for all $i \in \{1, \dots, n\}$, $Y_i(u) = 0$ and $S_n(u, \hat{\beta}) = 0$. So, we can rewrite for all $h \in \mathcal{H}_n$ that

$$\hat{\alpha}_h^{\hat{\beta}}(t) - \bar{\alpha}_h(t) = \frac{1}{nh} \sum_{i=1}^n \int_0^\tau K\left(\frac{t-u}{h}\right) \frac{S(u, \beta_0) - S_n(u, \hat{\beta})}{S_n(u, \hat{\beta}) S(u, \beta_0)} \mathbb{1}_{\{\bar{Y}(u) > 0\}} dN_i(u). \quad (29)$$

Consider the following sets:

$$\Omega_{H,k} = \left\{ \omega : \forall u \in [0, \tau], |S_n(u, \hat{\beta}) - S(u, \beta_0)| \leq 2C(s) B e^{BR} e^{2B|\beta_0|_1} \sqrt{\frac{\log(pn^k)}{n}} \right\}, \quad (30)$$

$$\Omega_{S_n} = \left\{ \omega : \forall u \in [0, \tau], S_n(u, \hat{\beta}) - S(u, \beta_0) \geq -\frac{c_S}{2} \right\}, \quad (31)$$

$$\Omega_k = \Omega_{H,k} \cap \Omega_{S_n}. \quad (32)$$

We decompose T_3 on Ω_k and on its complement. On Ω_k^c , let introduce the following lemma:

Lemma 6.6. *Under Assumptions 3.4.(ii)-(iii), 3.5.(i)-(iii), 3.1 and 3.2, for all $k \in \mathbb{N}$, we have*

$$\mathbb{E}[\|\hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h\|_2^2 \mathbb{1}(\Omega_k^c)] \leq c_3 n^{4-k/2},$$

where c_3 is a constant depending on B , $|\beta_0|_1$, R , $\|\alpha_0\|_{\infty, \tau}$, c_S , τ , $\|K\|_{\infty}$. Choosing $k \geq 10$ yields $\mathbb{E}[\|\hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h\|_2^2] \leq c_3/n$.

From Lemma 6.6,

$$\begin{aligned}\mathbb{E}\left[\sup_{h' \in \mathcal{H}_n} \|\hat{\alpha}_{h'}^{\hat{\beta}} - \bar{\alpha}_{h'}\|_2^2 \mathbf{1}(\Omega_k^c)\right] &\leq \sum_{h' \in \mathcal{H}_n} \mathbb{E}[\|\hat{\alpha}_{h'}^{\hat{\beta}} - \bar{\alpha}_{h'}\|_2^2 \mathbf{1}(\Omega_k^c)] \\ &\leq \sum_{h' \in \mathcal{H}_n} c_3 n^{4-k/2},\end{aligned}$$

which is of order $1/n$ as long as $k \geq 12$. On the other hand, from (29) on Ω_k , we have

$$\begin{aligned}\mathbb{E}\left[\sup_{h' \in \mathcal{H}_n} \int_0^\tau (\hat{\alpha}_{h'}^{\hat{\beta}} - \bar{\alpha}_{h'})^2(t) \mathbf{1}(\Omega_k) dt\right] \\ \leq \frac{16C(s)^2 B^2 e^{2BR} e^{4B|\beta_0|_1} \log(pn^k)}{c_S^2} \frac{1}{n} \mathbb{E}\left[\sup_{h' \in \mathcal{H}_n} \int_0^\tau \left(\int_0^\tau \frac{|K_{h'}(t-u)|}{S(u, \beta_0)} \left(\frac{1}{n} \sum_{i=1}^n dN_i(u)\right)\right)^2 dt\right].\end{aligned}$$

Then, we decompose $N_i = (N_i - \Lambda_i) + \Lambda_i$ to obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{h' \in \mathcal{H}_n} \int_0^\tau \left\{\int_0^\tau \frac{|K_{h'}(t-u)|}{S(u, \beta_0)} \left(\frac{1}{n} \sum_{i=1}^n dN_i(u)\right)\right\}^2 dt\right] \\ \leq 2\mathbb{E}\left[\sup_{h' \in \mathcal{H}_n} \int_0^\tau \left\{\int_0^\tau \frac{|K_{h'}(t-u)|}{S(u, \beta_0)} \left(\frac{1}{n} \sum_{i=1}^n dN_i(u) - \alpha_0(u)S(u, \beta_0)du\right)\right\}^2 dt\right] \quad (33)\end{aligned}$$

$$+ 2 \sup_{h' \in \mathcal{H}_n} \int_0^\tau \left\{\int_0^\tau |K_{h'}(t-u)| \alpha_0(u) du\right\}^2 dt. \quad (34)$$

The term (34) is bounded by $2\tau \|\alpha_0\|_{\infty, \tau}^2 \|K\|_{\mathbb{L}^1(\mathbb{R})}^2$. Let us bound the term (33),

$$\begin{aligned}\mathbb{E}\left[\sup_{h' \in \mathcal{H}_n} \int_0^\tau \left\{\int_0^\tau \frac{|K_{h'}(t-u)|}{S(u, \beta_0)} \left(\frac{1}{n} \sum_{i=1}^n dN_i(u) - \alpha_0(u)S(u, \beta_0)du\right)\right\}^2 dt\right] \\ \leq \sum_{h' \in \mathcal{H}_n} \int_0^\tau \text{Var}\left[\int_0^\tau \frac{|K_{h'}(t-u)|}{S(u, \beta_0)} \frac{1}{n} \sum_{i=1}^n dN_i(u)\right]\end{aligned}$$

It remains to bound the variance term.

$$\text{Var}\left[\frac{1}{n} \sum_{i=1}^n \int_0^\tau \frac{|K_h(t-u)|}{S(u, \beta_0)} dN_i(u)\right] \leq \frac{1}{n} \mathbb{E}\left[\left(\int_0^\tau \frac{|K_h(t-u)|}{S(u, \beta_0)} dN_1(u)\right)^2\right].$$

We apply the Doob-Meyer decomposition to get

$$\text{Var}\left[\frac{1}{n} \sum_{i=1}^n \int_0^\tau \frac{|K_h(t-u)|}{S(u, \beta_0)} dN_i(u)\right] \leq \frac{2}{n} \mathbb{E}\left[\left(\int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} dM_1(u)\right)^2\right] \quad (35)$$

$$+ \frac{2}{n} \mathbb{E}\left[\left(\int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \alpha_0(u) e^{\beta_0^T \mathbf{Z}_1} Y_1(u) du\right)^2\right]. \quad (36)$$

The term (35) is bounded by

$$\frac{2}{n} \mathbb{E} \left[\int_0^\tau \frac{K_{h'}^2(t-u)}{(S(u, \beta_0))^2} \alpha_0(u) e^{\beta_0^T \mathbf{Z}_1} Y_1(u) du \right] \leq \frac{2}{n} \frac{\|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}]}{c_S^2} \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{h'}, \quad (37)$$

and from the Cauchy-Schwarz inequality, the term (36) is bounded by

$$\frac{2}{n} \frac{\|\alpha_0\|_{\infty, \tau}^2 \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]_\tau}{c_S^2} \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{h'}. \quad (38)$$

From (37) and (38), (33) is bounded by

$$\frac{4}{n} \frac{\|\alpha_0\|_{\infty, \tau}}{c_S^2} \left[\mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] + \|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}] \right] \|K\|_{\mathbb{L}^2(\mathbb{R})}^2 \sum_{h' \in \mathcal{H}_n} \frac{1}{nh'}. \quad (39)$$

From Condition 3.5.(ii) and bounds (36) and (39), we deduce that

$$\begin{aligned} & \mathbb{E} \left[\sup_{h' \in \mathcal{H}_n} \int_0^\tau (\hat{\alpha}_{h'}^\beta - \bar{\alpha}_{h'})^2(t) \mathbf{1}(\Omega_k) dt \right] \\ & \leq C(s, c_S, B, R, |\beta_0|_1, \|\alpha_0\|_{\infty, \tau}, \tau, \|K\|_{\mathbb{L}^2(\mathbb{R})}, \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}], \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]) \frac{\log^a(n) \log(pn^k)}{n}. \end{aligned}$$

Finally, there exists a constant $c_5 > 0$ such that

$$\mathbb{E}[T_3] \leq c_5 \frac{\log^a(n) \log(n^k p)}{n}, \quad (40)$$

where c_5 depends on $s, c_S, B, R, \tau, \|\alpha_0\|_{\infty, \tau}, |\beta_0|_1, \|K\|_{\mathbb{L}^2(\mathbb{R})}, \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}]$ and $\mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]$.

- Study of $\mathbb{E}[T_4]$: Since

$$\hat{\alpha}_{h, h'}^\beta - \bar{\alpha}_{h, h'} = K_{h'} * (\hat{\alpha}_h^\beta - \bar{\alpha}_h),$$

we have from Young Inequality (Lemma 2.2 in the supplementary material) with $p = 1, q = 2, r = 2$,

$$\mathbb{E}[T_4] \leq \|K\|_{\mathbb{L}^1(\mathbb{R})}^2 \mathbb{E}[\|\hat{\alpha}_h^\beta - \bar{\alpha}_h\|_2^2] \leq C(s) \|K\|_{\mathbb{L}^1(\mathbb{R})}^2 \frac{\log(n^k p)}{n}, \quad (41)$$

where the last inequality is obtained from Lemma 6.3.

- Study of $\mathbb{E}[T_5]$: From Young Inequality (Lemma 2.2 in the supplementary material) with $p = 1, q = 2, r = 2$, we obtain that

$$\|\alpha_{h'} - \alpha_{h, h'}\|_2^2 = \|K_{h'} * (\alpha_0 - K_h * \alpha_0)\|_2^2 \leq \|K\|_{\mathbb{L}^1(\mathbb{R})}^2 \|\alpha_0 - K_h * \alpha_0\|_2^2$$

Therefore, since $\alpha_h = K_h * \alpha_0$,

$$\mathbb{E}[T_5] \leq \|K\|_{\mathbb{L}^1(\mathbb{R})}^2 \|\alpha_0 - \alpha_h\|_2^2, \quad (42)$$

which corresponds to a bias term.

Finally, gathering the bounds of the five terms (27), (28), (40), (41) and (42), gives the result of Proposition 6.4. \square

6.4 Proof of Lemma 6.5

We have to control the supremum of $\nu_{n,h}(f)$ defined by (43) over the ball $\mathcal{B}_\tau = \{g \in \mathbb{L}^2([0, \tau]), \|g\|_2 = 1\}$. For all $h \in \mathcal{H}_n$ and $f \in \mathcal{B}_\tau$, we have

$$\nu_{n,h}(f) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau f(t) \left(\int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} (dN_i(u) - \alpha_0(u)S(u, \beta_0)du) \right) dt. \quad (43)$$

Usually, to control such a process, we apply the Talagrand Inequality given in Theorem ?? . However, since $\nu_{n,h}(f)$ is not bounded, we can not directly apply the Talagrand Inequality: we have to introduce a truncation (see Chagny (2014) for a close approach). Let us define for a constant c ,

$$\kappa_n = c \frac{\sqrt{n}}{\log n},$$

and we decompose $\nu_{n,h}$ as

$$\nu_{n,h}(f) = \nu_{n,h}^{(1)}(f) + \nu_{n,h}^{(2)}(f),$$

where

$$\begin{aligned} \nu_{n,h}^{(1)}(f) &= \frac{1}{n} \sum_{i=1}^n \int_0^\tau f(t) \int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \mathbb{1}_{\{N_i(\tau) \leq \kappa_n\}} dN_i(u) dt \\ &\quad - \frac{1}{n} \sum_{i=1}^n \int_0^\tau f(t) \int_0^\tau \mathbb{E} \left[\frac{K_h(t-u)}{S(u, \beta_0)} \mathbb{1}_{\{N_i(\tau) \leq \kappa_n\}} dN_i(u) \right] dt, \end{aligned}$$

and

$$\begin{aligned} \nu_{n,h}^{(2)}(f) &= \frac{1}{n} \sum_{i=1}^n \int_0^\tau f(t) \int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \mathbb{1}_{\{N_i(\tau) > \kappa_n\}} dN_i(u) dt \\ &\quad - \frac{1}{n} \sum_{i=1}^n \int_h^{\tau-h} f(t) \int_0^\tau \mathbb{E} \left[\frac{K_h(t-u)}{S(u, \beta_0)} \mathbb{1}_{\{N_i(\tau) > \kappa_n\}} dN_i(u) \right] dt. \end{aligned}$$

- Control of $\nu_{n,h}^{(1)}(f)$:

We can apply a Talagrand Inequality to $\nu_{n,h}^{(1)}(f)$, which is bounded. To apply this concentration inequality, we need to determine the bounds H , M , W and the constant ε (see Theorem ?? in Appendix ?? for the notations).

- Determination of the constant M :

Using the Cauchy-Schwarz Inequality, we have for

$$\begin{aligned}
& \left| \int_0^\tau f(t) \int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \mathbf{1}_{\{N_1(\tau) \leq \kappa_n\}} dN_1(u) dt \right| \\
& \leq \|f\|_2 \left| \int_0^\tau \left(\int_0^\tau K_h^2(t-u) dt \right)^{1/2} \frac{\mathbf{1}_{\{N_1(\tau) \leq \kappa_n\}}}{S(u, \beta_0)} dN_1(u) \right| \\
& \leq \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{\sqrt{h}} \frac{|N_1(\tau) \mathbf{1}_{\{N_1(\tau) \leq \kappa_n\}}|}{c_S} \\
& \leq \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{c_S \sqrt{h}} \kappa_n := M \sim \frac{\sqrt{n}}{\log n \sqrt{h}}.
\end{aligned}$$

– Determination of the constant H :

Let define

$$\psi_h(t) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \mathbf{1}_{\{N_i(\tau) \leq \kappa_n\}} dN_i(u)$$

We have $\sup_{f \in \mathcal{B}_\tau} (\nu_{n,h}^{(1)}(f))^2 = \sup_{f \in \mathcal{B}_\tau} \langle \psi_h - \mathbb{E}[\psi_h], f \rangle_2^2 = \|\psi_h - \mathbb{E}[\psi_h]\|_2^2$. We deduce from the Doob-Meier decomposition that

$$\begin{aligned}
\mathbb{E} \left[\sup_{f \in \mathcal{B}_\tau} (\nu_{n,h}^{(1)}(f))^2 \right] &= \int_0^\tau \text{Var}[\psi_h(t)] dt \\
&\leq \frac{1}{n} \int_0^\tau \mathbb{E} \left[\left(\int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \mathbf{1}_{\{N_1(\tau) \leq \kappa_n\}} dN_1(u) \right)^2 \right] dt \\
&\leq \frac{2\|\alpha_0\|_{\infty, \tau}^2}{c_S^2} \left(\|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]_\tau + \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] \right) \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh} := H^2
\end{aligned}$$

We have $H^2 = V(h)/\kappa$. Then, we set $\varepsilon^2 = 1/2$ and $\kappa = 80$ in order to have $2(1 + 2\varepsilon^2)H^2 = V(h)/20 = O(1/nh)$.

– Determination of the constant W :

Since $f \in \mathcal{B}_\tau$, we have

$$\begin{aligned}
& \text{Var} \left[\int_0^\tau f(t) \int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \mathbf{1}_{\{N_1(\tau) \leq \kappa_n\}} dN_1(u) dt \right] \\
& \leq \mathbb{E} \left[\left(\int_0^\tau f(t) \int_0^\tau \frac{K_h(t-u)}{S(u, \beta_0)} \mathbf{1}_{\{N_1(\tau) \leq \kappa_n\}} dN_1(u) dt \right)^2 \right] \\
& \leq \mathbb{E} \left[\mathbf{1}_{\{N_1(\tau) \leq \kappa_n\}} \left(\int_0^\tau \frac{(K_h * f)(u)}{S(u, \beta_0)} dN_1(u) \right)^2 \right].
\end{aligned}$$

So, from the Doob-Meier decomposition and Young Lemma 2.2 in the supplementary material, we have

$$\begin{aligned}
\text{Var} \left[\int_0^\tau f(t) \int_0^\tau \frac{K_h(t-u)}{S(u, \boldsymbol{\beta}_0)} \mathbb{1}_{\{N_1(\tau) \leq \kappa_n\}} dN_1(u) dt \right] \\
\leq \frac{2\|\alpha_0\|_{\infty, \tau}}{c_S^2} \left[\tau \|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}] + \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] \right] \|K_h * f\|_2^2 \\
\leq \frac{2\|\alpha_0\|_{\infty, \tau}}{c_S^2} \left[\tau \|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}] + \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] \right] \|K\|_{\mathbb{L}^1(\mathbb{R})}^2 := W
\end{aligned}$$

Then, from Assumptions 3.7.(ii) and (iii), we deduce that

$$\begin{aligned}
\sum_{h \in \mathcal{H}_n} \mathbb{E} \left[\left\{ \sup_{f \in \mathcal{B}_\tau} (\nu_{n,h}^{(1)}(f))^2 - V(h)/20 \right\}_+ \right] &\leq \frac{\vartheta_1}{n} \sum_{h \in \mathcal{H}_n} \left(e^{-\frac{\vartheta_2}{h}} + \frac{1}{n \log^2(n)h} e^{-\vartheta_3 \log n} \right) \\
&\leq \frac{\tilde{\vartheta}_1}{n} + \tilde{\vartheta}_2 \frac{\log^{a-2} n}{n^{\tilde{\vartheta}_3}} \tag{44}
\end{aligned}$$

with

$$V(h) = \kappa \frac{2\|\alpha_0\|_{\infty, \tau}}{c_S^2} \left(\|\alpha_0\|_{\infty, \tau} \mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]_\tau + \mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}] \right) \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nh}.$$

- Control of $\nu_{n,h}^{(2)}(f)$:

Now, let us focus on the second unbounded term $\nu_{n,h}^{(2)}(f)$. Let us consider the process $\Psi(t)$ defined as

$$\frac{1}{n} \sum_{i=1}^n \left[\int_0^\tau \frac{K_h(t-u)}{S(u, \boldsymbol{\beta}_0)} \mathbb{1}_{\{N_i(\tau) > \kappa_n\}} dN_i(u) - \mathbb{E} \left[\int_0^\tau \frac{K_h(t-u)}{S(u, \boldsymbol{\beta}_0)} \mathbb{1}_{\{N_i(\tau) > \kappa_n\}} dN_i(u) \right] \right],$$

so that $\nu_{n,h}^{(2)}(f) = \int_0^\tau f(t) \Psi(t) dt$. Using Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\mathbb{E} \left[\sup_{f \in \mathcal{B}_\tau} (\nu_{n,h}^{(2)}(f))^2 \right] &\leq \mathbb{E} \left[\int_0^\tau \Psi^2(t) dt \right] \\
&\leq \frac{1}{n} \int_0^\tau \text{Var} \left[\int_0^\tau \frac{K_h(t-u)}{S(u, \boldsymbol{\beta}_0)} \mathbb{1}_{\{N_1(\tau) > \kappa_n\}} dN_1(u) \right] dt \\
&\leq \frac{1}{n} \int_0^\tau \mathbb{E} \left[\left(\int_0^\tau \frac{K_h(t-u)}{S(u, \boldsymbol{\beta}_0)} \mathbb{1}_{\{N_1(\tau) > \kappa_n\}} dN_1(u) \right)^2 \right] dt
\end{aligned}$$

Applying the Cauchy-Schwarz Inequality (see Lemma 2.1 in the supplementary ma-

terial), we obtain that for all $k > 0$,

$$\begin{aligned}
\mathbb{E} \left[\sup_{f \in \mathcal{B}_\tau} (\nu_{n,h}^{(2)}(f))^2 \right] &\leq \frac{1}{n} \int_0^\tau \mathbb{E} \left[\mathbb{1}_{\{N_1(\tau) > \kappa_n\}} N_1(\tau) \int_0^\tau \frac{K_h^2(t-u)}{S^2(u, \beta_0)} dN_1(u) \right] dt \\
&\leq \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nhc_S^2} \mathbb{E}[N_1^2(\tau) \mathbb{1}_{\{N_1(\tau) > \kappa_n\}}] \\
&\leq \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nhc_S^2} \frac{\mathbb{E}[N_1^{k+2}(\tau)]}{\kappa_n^k} \\
&\leq \frac{\|K\|_{\mathbb{L}^2(\mathbb{R})}^2}{nhc_S^2} \frac{\mathbb{E}[N_1^{k+2}(\tau)]}{n}
\end{aligned}$$

From Assumption 3.7.(ii), we deduce that for k large enough

$$\sum_{h \in \mathcal{H}_n} \mathbb{E} \left[\sup_{f \in \mathcal{B}_\tau} (\nu_{n,h}^{(2)}(f))^2 \right] \leq C \frac{\log^a(n) \mathbb{E}[N(\tau)^{k+1}]}{n}.$$

It remains to verify that $\mathbb{E}[N(\tau)^{k+1}]$ is bounded. Using the fact that for all $a \geq 0$, $b \geq 0$ and $p \geq 1$, $(a+b)^k \leq 2^{k-1}(a^k + b^k)$ and from the B rkholder Inequality, we can easily show by recurrence that for all $p \in \mathbb{N}^*$, $\mathbb{E}[N(\tau)^k] \leq C_k$. Thus, we conclude that for a good choice of p ,

$$\sum_{h \in \mathcal{H}_n} \mathbb{E} \left[\sup_{f \in \mathcal{B}_\tau} (\nu_{n,h}^{(2)}(f))^2 \right] \leq \tilde{C} \frac{\log^a(n)}{n}, \tag{45}$$

for a constant $\tilde{C} > 0$.

Combining (44) and (45), we finally get

$$\sum_{h \in \mathcal{H}_n} \mathbb{E} \left[\left\{ \sup_{f \in \mathcal{B}_\tau} \nu_{n,h}^2(f) - V(h)/10 \right\}_+ \right] \leq \frac{c_6}{n} + c_7 \frac{\log^a(n)}{n},$$

where c_6 and c_7 depends on τ , $\|\alpha_0\|_{\infty, \tau}$, c_S , $\mathbb{E}[e^{\beta_0^T \mathbf{Z}_1}]$, $\mathbb{E}[e^{2\beta_0^T \mathbf{Z}_1}]$, $\|K\|_{\mathbb{L}^1(\mathbb{R})}$ and $\|K\|_{\mathbb{L}^2(\mathbb{R})}$. \square

Remark: A similar lemma can be obtained for the centered process $\langle \bar{\alpha}_{h,h'} - \alpha_{h,h'}, f \rangle_2$, where $\bar{\alpha}_{h,h'} = K_{h'} * \bar{\alpha}_h$ and $\alpha_{h,h'} = \mathbb{E}[\bar{\alpha}_{h,h'}]$ for $h, h' \in \mathcal{H}_n$. Indeed, from Young Lemma 2.2 in the supplementary material, we have

$$\begin{aligned}
\langle \bar{\alpha}_{h,h'} - \alpha_{h,h'}, f \rangle_2 &= \int_0^\tau f(t) \left(K_{h'} * \bar{\alpha}_h(t) - \mathbb{E}[K_{h'} * \bar{\alpha}_h(t)] \right) dt \\
&\leq \|f\|_2 \|K_{h'} * (\bar{\alpha}_h - \mathbb{E}[\bar{\alpha}_h])\|_2 \\
&\leq \|f\|_2 \|K\|_{\mathbb{L}^1(\mathbb{R})} \|\bar{\alpha}_h - \mathbb{E}[\bar{\alpha}_h]\|_2^2.
\end{aligned}$$

Just take the same constants M , H^2 and W than previously and multiply them by $\|K\|_{\mathbb{L}^1(\mathbb{R})}$.

The proofs of Lemmas 6.2, 6.3 and 6.6 are available in the supplementary material.

References

- O. Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wista, 1978)*, volume 2 of *Lecture Notes in Statist.*, pages 1–25. Springer, New York, 1980.
- P. K. Andersen, Ø. Borgan, R. D. Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0.
- O. Bouaziz, F. Comte, and A. Guillaux. Nonparametric estimation of the intensity function of a recurrent event process. *Statistica Sinica*, 23(2):635–665, 2013.
- J. Bradic, Fan, J., and J. Jiang. Regularization for Cox’s proportional hazards model with NP-dimensionality. *The Annals of Statistics*, 39(6):pp. 3092–3120, 2012.
- J. Bradic and R. Song. Gaussian Oracle Inequalities for Structured Selection in Non-Parametric Cox Model. *arXiv preprint arXiv:1207.4510*, 2012.
- G. Chagny. *Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie*. PhD thesis, Université René Descartes-Paris V, 2013.
- G. Chagny. Adaptive warped kernel estimators. *accepted for publication in Scandinavian Journal of statistics*, 2014.
- D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B. (Methodological)*, 34:pp. 187–220, 1972. ISSN 0035-9246.
- M. Doumic, M. Hoffmann, P. Reynaud-Bouret, and V. Rivoirard. Nonparametric estimation of the division rate of a size-structured population. *SIAM Journal on Numerical Analysis*, 50(2):925–950, 2012.
- J. Fan, Y. Feng, and Y. Wu. High-dimensional variable selection for Cox’s proportional hazards model. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*, pages 70–86. Institute of Mathematical Statistics, 2010.
- T.R. Fleming and D.P. Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.
- A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.
- G. Grégoire. Least squares cross-validation for counting process intensities. *Scandinavian journal of statistics*, pages 343–360, 1993.

- A. Guillaoux, S. Lemler, and M.L. Taupin. Adaptive estimation of the baseline hazard function in the cox model by model selection, with high-dimensional covariates. *arXiv preprint arXiv:1503.00226*, 2015.
- J. Huang, T. Sun, Z. Ying, Y. Yu, and C.H. Zhang. Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics*, 41(3):1142–1165, 2013.
- S. Kong and B. Nan. Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Arxiv preprint arXiv:1204.1992*, 2012.
- S. Lemler. *Estimation for counting processes with high-dimensional covariates*. PhD thesis, Université d’Évry Val d’Essonne, 2014.
- S. Loi, B. Haihe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J.A. Foekens, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of clinical oncology*, 25(10):1239–1246, 2007.
- J.S. Marron and W.J. Padgett. Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics*, pages 1520–1535, 1987.
- H. Ramlau-Hansen. *Udglatning med kernefunktioner i forbindelse med tælleprocesser: Del 1*. Forsikringsmatematisk Laboratorium, Københavns universitet, 1981.
- H. Ramlau-Hansen. The choice of a kernel function in the graduation of counting process intensities. *Scandinavian Actuarial Journal*, 1983(3):165–182, 1983a.
- H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pages 453–466, 1983b.
- L. Tian, A. Alizadeh, A. Gentles, and R. Tibshirani. A simple method for detecting interactions between a treatment and a large number of covariates. *arXiv preprint arXiv:1212.2995*, 2012.
- S. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):pp. 614–645, 2008.